Data Scientist Challenge - LATAM Airlines

Instrucciones

En Advanced Analytics valoramos muchísimo el trabajo en equipo y la constante interacción entre los distintos roles que trabajan en un producto basado en datos, como el Data Scientist, Machine Learning Engineering, Data Engineer, entre otros. Es por este motivo que una habilidad esencial que buscamos a la hora de buscar nuevos integrantes es el manejo adecuado de git. **Este desafío deberá ser entregado en la plataforma de git que más te acomode y que sea pública para que la podamos revisar.** Lo que buscamos con esto es poder entender de mejor manera el desarrollo que generaste con tu código, cómo lo fuiste mejorando en el tiempo y si tienes proyectos propios en este repositorio nos servirán para conocer mejor tu experiencia en base a tu propio portafolio.

Instrucciones Git:

1) Crear un repositorio en la plataforma de git que más te acomode y que sea pública
2) Haber trabajado con una rama principal y otra de desarrollo. Opcional, ocupar alguna práctica de desarrollo de GitFlow.

Instrucciones del desafío:

1) Debes enviar el link al repositorio al mail cristobal.guzman@latam.com con asunto `Challenge Data Scientist - [Nombre][Apellido]`, ejemplo `Challenge Data Scientist - Cristobal Guzman`.
2) Se aceptará los cambios en el repositorio hasta el **Lunes 14 de Marzo de 2022 a las 23:59 hrs**.
3) En la siguiente carpeta de Google Drive encontrarás las instrucciones del desafío y el archivo `dataset_SCL.csv` que utilizarás para desarrollarlo.
4) El repositorio debe tener un jupyter notebook llamado `solution.ipynb` utilizando python 3. **No serán revisados otros lenguajes como R o similar**.
5) En `solution.ipynb` deben estar resueltas las respuestas a todas las preguntas del desafío
6) Dentro del repositorio deben estar todos los archivos necesarios para que los evaluadores puedan clonar y luego correr tu notebook sin problemas
7) Una copia de tu CV (curriculum vitae) en formato .pdf

Problema

El problema consiste en predecir la probabilidad de atraso de los vuelos que aterrizan o despegan del aeropuerto de Santiago de Chile (SCL). Para eso les entregamos un dataset usando datos públicos y reales donde cada fila corresponde a un vuelo que aterrizó o despegó de SCL. Para cada vuelo se cuenta con la siguiente información:

Fecha-I : Fecha y hora programada del vuelo.
Vlo-I : Número de vuelo programado.
Ori-I : Código de ciudad de origen programado.
Des-I : Código de ciudad de destino programado.
Emp-I : Código aerolínea de vuelo programado.
Fecha-O : Fecha y hora de operación del vuelo.
Vlo-O : Número de vuelo de operación del vuelo.
Ori-O : Código de ciudad de origen de operación
Des-O : Código de ciudad de destino de operación.
Emp-O : Código aerolínea de vuelo operado.
DIA : Día del mes de operación del vuelo.
MES : Número de mes de operación del vuelo.
AÑO : Año de operación del vuelo.
DIANOM : Día de la semana de operación del vuelo.
TIPOVUELO : Tipo de vuelo, I =Internacional, N =Nacional.
OPERA : Nombre de aerolínea que opera.
SIGLAORI : Nombre ciudad origen.
SIGLADES : Nombre ciudad destino.

Desafío

1. ¿Cómo se distribuyen los datos? ¿Qué te llama la atención o cuál es tu conclusión sobre esto?
2. Genera las columnas adicionales y luego expórtelas en un archivo `synthetic_features.csv` :
    - `temporada_alta` : 1 si Fecha-I está entre 15-Dic y 3-Mar, o 15-Jul y 31-Jul, o 11-Sep y 30-Sep, 0 si no.
    - `dif_min` : diferencia en minutos entre Fecha-O y Fecha-I .
    - `atraso_15` : 1 si dif_min > 15, 0 si no.
    - `periodo_dia` : mañana (entre 5:00 y 11:59), tarde (entre 12:00 y 18:59) y noche (entre 19:00 y 4:59), en base a Fecha-I .
3. ¿Cómo se compone la tasa de atraso por destino, aerolínea, mes del año, día de la semana, temporada, tipo de vuelo? ¿Qué variables esperarías que más influyeran en predecir atrasos?
4. Entrena uno o varios modelos (usando el/los algoritmo(s) que prefieras) para estimar la probabilidad de atraso de un vuelo. Siéntete libre de generar variables adicionales y/o complementar con variables externas.
5. Evalúa tu modelo. ¿Qué performance tiene? ¿Qué métricas usaste para evaluar esa performance y por qué? ¿Por qué elegiste ese algoritmo en particular? ¿Qué variables son las que más influyen en la predicción? ¿Cómo podrías mejorar la performance?

Aspectos a considerar

Orden y claridad al momento de plantear un análisis, idea, código, etc.
Creatividad para resolver el desafío.
Código versionado en Git.
No vamos a revisar excel, macros, códigos en R.
No vamos a revisar desafíos que no lleguen en la fecha indicada
Ante cualquier duda, deja explícitos tus supestos
No vivimos en tu cabeza, trata de expresarte lo mejor posible para explicar tus decisiones y respuestas